MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2019/2020

TDS3301 – DATA MINING

(All sections / Groups)

10 MARCH 2020
9:00 a.m. – 11:00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENT

1. This question paper consists of **5 pages** including cover page with **4 questions** only.
2. Answer **ALL** questions. All questions carry equal marks and the distributions of marks for each question is given.
3. Please write all your answers in the **Answer Booklet** provided.

Question 1

Table 1: Student information from a university database

Name	Age (years)	Height (cm)	Weight (kg)	Gender	Hair color	Current Grade	Study Year
Ash	24	165	70	Male	Black	A	2
Brock	20	174	76	Male	Brown	C	1
Cilan	22	180	73	Male		A+	3
Dawn	21	158		Female	Brown	B	2
Ethan	24	178	80	Male	Black	B-	1
Fred	24		83	Male	Black	B	1
Gary		178	74	Male	Black	A	2
Holly	21	175	60	Female		A-	2
Isabelle	22	165	65	Female	Brown	C	3
Jacob	22		77		Black	C+	3
Kenny	22	165		Male	Black	A+	2
Lionel	23	170	86	Male		A	1
Monica		158	73	Female	Black	B	2
Nicole	20	157	55		Black	B	3
Oliver	21	180	73	Male	Brown	B-	4
Pablo	21	173	59	Male	Brown	B+	1
Quill	21	170	73	Male		C	4
Rufus	23	161	77	Male	Black	A	2
Scott	23	171	69	Male	Black	C-	3

(a) Identify the attribute types of **ALL** the columns in *Table 1.* (4 marks)

(b) List **FOUR** possible reasons that caused the *incomplete* data of *Table 1.* (2 marks)

(c) The data for *Age, Height,* and *Weight* can be cleaned using basic statistical description methods. Suggest one method that can be used and explain your choice. (1 marks)

(d) Using the method suggested in *Question 1 (c)*, clean the data for the *Age, Height,* and *Weight* attributes **ONLY**. Show your workings and show the cleaned data in a simplified table. Round your answers to **ZERO** decimal points. (3 marks)
Example table:

Name	Age (years)	Height (cm)	Weight (kg)
Dawn	21	158	
Fred	24		83
Gary		178	74
Jacob	22		77
Kenny	22	165	
Monica		158	73

[TOTAL 10 MARKS]

Continued ...

Question 2

Table 2: Record of taekwondo moves by a taekwondo student

Competitions	Moves
0701	Forefist, Elbow strike, Front kick, Jump kick
2301	Forefist, Front kick, Side kick, Back kick, Axe kick
0803	Side kick, Forefist, Elbow strike, Front kick
0205	Forefist, Front kick, Elbow strike, Back kick, Axe kick
1608	Knifehand, Side kick, Elbow strike, Front kick

(a) *Table 2* shows a taekwondo student's moves in multiple competitions as a transactional record. At the minimum support of 60% and minimum confidence of 80%, find the frequent itemsets using the *FP-growth* method. (6 marks)

(b) Find all the strong association rules that satisfy the minimum support and confidence thresholds. Show your workings in a table and list down the strong rules using the $X \rightarrow Y(s, c)$ format. (3.5 marks)

Example table:

Association	Support (%)	Confidence (%)
$X \rightarrow Y$		
$Y \rightarrow X$		

(c) Suggest a possible usage of the patterns found in *Question 2 (b)* to a person who is going to have a match with the taekwondo student recorded in *Table 2.*

(0.5 marks)

Formulae:

Support, $s(X \rightarrow Y) = \frac{X \cup Y}{N}$

Confidence, $c(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$

[TOTAL 10 MARKS]

Continued ...

Question 3

Table 3: Vehicle status records from the Road and Transport Department

Car plate	Price (RM)	Type	Origin	Stolen?
AX1234	101k-200k	Sports	Domestic	Y
BPX1235	71k-100k	Sports	Domestic	N
CCU1236	101k-200k	Sports	Domestic	Y
DT1237	45k-70k	Sports	Domestic	N
EAT1238	71k-100k	Sports	Imported	Y
FX1239	45k-70k	SUV	Imported	N
HWA1240	45k-70k	SUV	Imported	Y
JJT1241	45k-70k	SUV	Domestic	N
KLM1242	101k-200k	SUV	Imported	N
MAA1243	101k-200k	Sports	Imported	Y

(a) Perform feature selection on the data given in *Table 3*, for a classification task that attempts to classify if a vehicle are at risk to be stolen. Give reason for your selection. (2 marks)

(b) Given a car with the following features:

Car plate	Price (RM)	Type	Origin	Stolen?
NSX8347	71k-100k	SUV	Imported	?

classify if this car is at risk using *Naïve Bayesian Classification.* Show workings. (6 marks)

(c) Explain the assumption used to perform the classification in *Question 3 (b)*. (1 mark)

(d) Describe the algorithm that eliminates the assumption chosen in *Question 3 (c)* to improve the performance of the *Naïve Bayesian Classification.* (1 mark)

Formulae:

$P(X|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times ... \times P(x_n|C_i)$

Maximum posteriori, $P(C_i|X) = P(X|C_i)P(C_i)$

[TOTAL 10 MARKS]

Continued...

Question 4

Consider the following scenario:
Company S is an online music streaming platform. The platform has an extensive database containing a large collection of music and miscellaneous information regarding the files. The users of this platform are from various countries around the world.

(a) State the type of database used by Company S to store their data. (1 mark)

(b) Company S intends to profile the users' music preferences based on their country of origin, i.e. people from which country prefers what type of music. State **FOUR** attributes which can be used for this task. (2 marks)

(c) Clustering is a method that can be used to complete the task stated in *Question 4 (b)*. Describe the steps of the *k-means* algorithm for this task. (4 marks)

(d) A new user from Country M has just signed up for the streaming platform. Describe how the outcome in *Question 4 (c)* can be used to generate a music playlist recommendation for this user. (1 marks)

(e) The current data of Company S is organized on a country level. The company now wish to examine the user patterns from a regional level, e.g. Southeast Asia, North America, etc. State and describe the data preprocessing step required for this task. (2 marks)

[TOTAL 10 MARKS]

End of page.